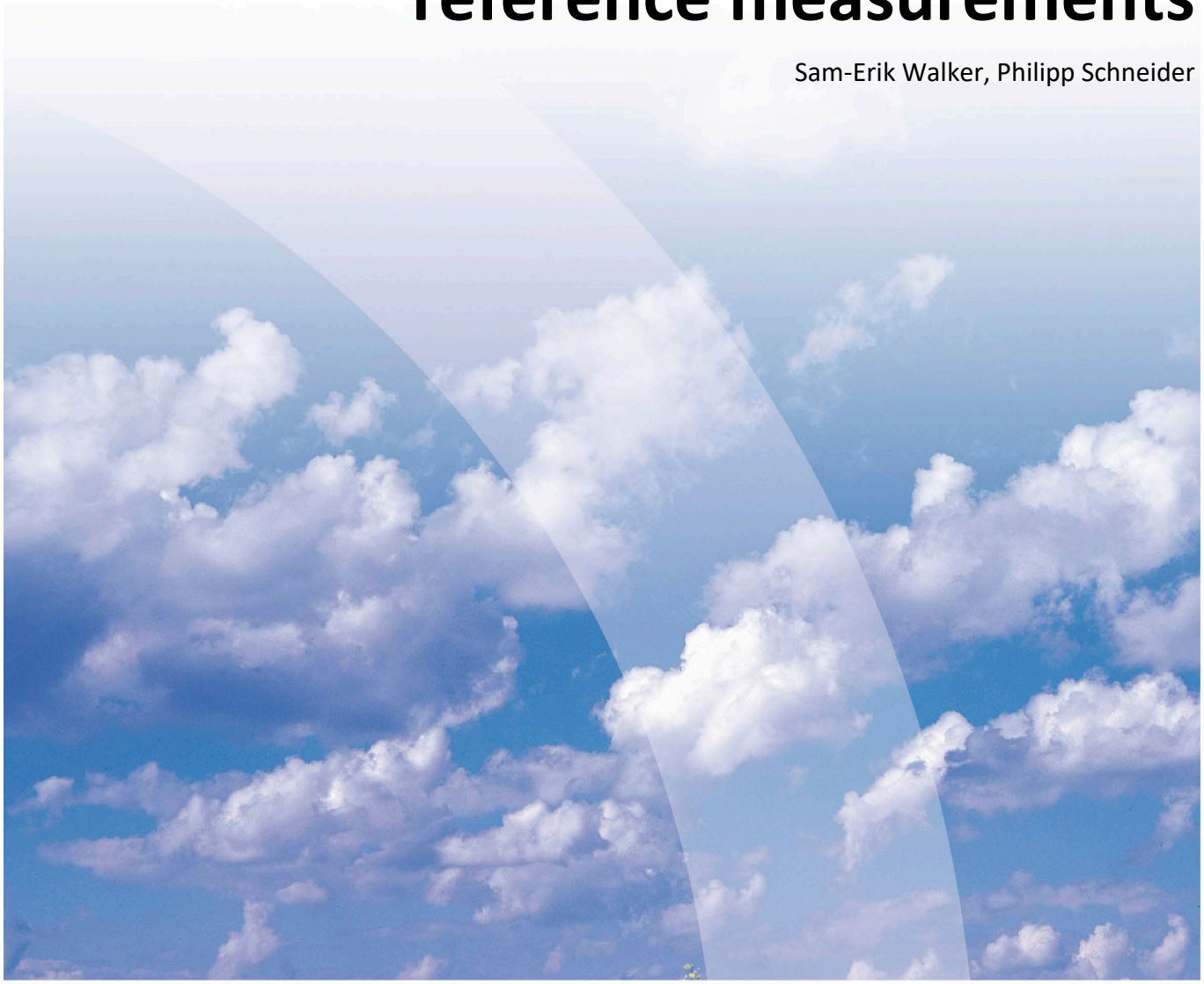
Norsk institutt for luftforskning
Norwegian Institute for Air Research

A study of the relative expanded uncertainty formula for comparing low-cost sensor and reference measurements

Sam-Erik Walker, Philipp Schneider



NILU report 01/2020

NILU report 01/2020	ISBN: 978-82-425-2997-8 ISSN: 2464-3327	CLASSIFICATION: A – Unclassified (open report)
DATE 29.04.2020	SIGNATURE OF RESPONSIBLE PERSON Ole-Anders Braathen, Deputy Director (sign.)	NUMBER OF PAGES 37

TITLE	PROJECT LEADER
A study of the relative expanded uncertainty formula for comparing low-cost sensor and reference measurements	Sam-Erik Walker
	NILU PROJECT NO. 119030
AUTHOR(S) Sam-Erik Walker, Philipp Schneider	QUALITY CONTROLLER Britt Ann Kåstad Høiskar
REPORT PREPARED FOR Norwegian Institute for Air Research (NILU), P.O. Box 100, NO-2027 Kjeller.	CONTRACT REF.

ABSTRACT

In this report, we investigate the relative expanded uncertainty (REU) formula for comparing low-cost sensors (microsensors) and reference measurements. The purpose of the REU formula is to check if microsensor measurements follow the data quality objective (DQO) of the European Air Quality Directive 2008/50/EC to be considered equivalent to a reference instrument. The project aimed to obtain a good understanding of the REU formula for its proper use in current and future projects involving microsensors.

NORWEGIAN TITLE

Undersøkelse av usikkerhetsformel for sammenligning av mikrosensorer med referansemålinger

KEYWORDS

Air quality | Environmental monitoring | Quality systems

ABSTRACT (in Norwegian)

Rapporten inneholder en undersøkelse av en usikkerhetsformel som brukes for å sammenligne målinger gjort med mikrosensorer med referansemålinger. Hensikten med formelen er å definere en relativ usikkerhet i målinger gjort med rimelige mikrosensorer for å sjekke om kvalitetsstandarden relatert til det europeiske luftkvalitetsdirektivet (2008/50/EC) overholdes for disse sensorene sammenlignet med referansemålinger.

PUBLICATION TYPE: Digital document (pdf)	COVER PICTURE: Source: NILU

Preface

This is the final report from an internal project at NILU to investigate the relative expanded uncertainty (REU) formula for comparing low-cost sensors (microsensors) and reference measurements. The purpose of the REU formula is to check if microsensor measurements follow the data quality objective (DQO) of the European Air Quality Directive 2008/50/EC to be considered equivalent to reference instruments. The project aimed to obtain a good understanding of the REU formula for its proper use in projects involving microsensors.

The work was led by Sam-Erik Walker and carried out in collaboration with Philipp Schneider and Jean-Marie Lepioufle. Thanks also to Matthias Vogt, Franck Dauge, and Leif Marsteen for valuable input.

Contents

Summary

The relative expanded uncertainty formula, as described in the EC Guidance document from 2010, is used to define uncertainties of low-cost sensor measurements relative to reference measurements to check if the Data Quality Objective of the European Air Quality Directive 2008/50/EC, primarily for indicative methods, is reached for these sensors as compared with the reference instrument.

In this report, we investigate the relative expanded uncertainty formula both theoretically and practically, to properly understand the background for and derivation of this formula for its proper use in current and future projects involving low-cost or microsensors. We show that this formula is correct under the usual assumptions of orthogonal regression, but that it also needs an additional assumption of an expected zero difference or bias between the measurands for each time point to represent a proper two times relative standard deviation value. The latter assumption is in practice equivalent to assuming an ideal 45-degree straight-line relationship between the low-cost and reference instrument measurands. If this assumption is not correct, a value calculated with this formula may still be used but must then be interpreted as an upper bound of the relative expanded uncertainty (REU) at the indicated low-cost measurement level.

The report stresses the importance of performing orthogonal regression in a proper way, not through the standard orthogonal regression formulae in the literature, but by applying the two-step adjusted orthogonal regression approach as developed by Dissanaike and Wang in their paper from 2003. We also show an alternative formula for the relative expanded uncertainty following more closely the model assumed for this two-step adjusted orthogonal regression method. The alternative formulae seem to give overall slightly higher relative expanded uncertainty values as compared with the standard method. Alternatively, one may use the standard relative uncertainty formula in combination with the two-step adjusted orthogonal regression method.

Simulation results show that the two-step adjusted orthogonal regression method works well in combination with both the standard and alternative relative expanded uncertainty formula.

A study of the relative expanded uncertainty formula for comparing low-cost sensor and reference measurements

1 Introduction

This project aims to gain better insight and an improved understanding of the relative expanded uncertainty (REU) formula as used in a recent paper by Spinelle et al. (2015) and defined in the EC Guide to the demonstration of equivalence document (EC, 2010). This formula is used to define relative uncertainties of low-cost sensors for checking if the Data Quality Objective (DQO) of the European Air Quality Directive (2008/50/EC) (EC, 2008) for indicative methods are reached for these sensors as compared with reference measurements.

The REU formula has been used by NILU in previous projects involving the calibration of low-cost sensors. It is therefore important to properly understand the background for and derivation of this formula for its proper use in current and future projects involving such sensors, e.g. in connection with the current IFLINK project. This is also important if we are going to use these sensors to compare with modelled concentrations or to combine with such concentrations.

The uncertainty referred to above is described as follows in the EC directive Annex I (EC, 2008):

"The uncertainty (expressed at a 95 % confidence level) of the assessment methods will be evaluated in accordance with the principles of the CEN Guide to the Expression of Uncertainty in Measurement (ENV 13005-1999), the methodology of ISO 5725:1994 and the guidance provided in the CEN report 'Air Quality — Approach to Uncertainty Estimation for Ambient Air Reference Measurement Methods' (CR 14377:2002E). The percentages for uncertainty in the above table are given for individual measurements averaged over the period considered by the limit value (or target value in the case of ozone), for a 95 % confidence interval."

The EC Guidance document (EC, 2010), which is a follow-up on the EC directive, describes that the REU of a measurement of an assessment method (i.e. low-cost sensor in our case) should, in general, be expressed by the following formula

$$U_r(y) = \frac{k\sqrt{V(y)}}{y}, \tag{1.1}$$

where y is the measured concentration; $V(y)$ is the estimated variance of the assessment method, i.e. the estimated variance of the low-cost sensor measurement in our case; and k is a given so-called coverage factor related to the confidence level associated with the uncertainty statement.

Since one generally assumes that measurement errors are normally distributed, a 95 % confidence level as stated in the EC directive Annex I quote and repeated in the EC Guidance document corresponds to using a factor $k = 2$ in the formula (1.1). Without this factor, i.e. for $k = 1$, the expression in (1.1) is defined simply as the relative uncertainty of the measurement y.

REUs expressed in the form of percentages rather than as fractions of 1 is obtained by multiplying the expression in (1.1) by 100. These may be compared with threshold percentages for assessment methods as stated in the table in Annex I part A of the EC directive (EC, 2008). The relevant percentages for various compounds for indicative measurement methods are reproduced in Table 1.

Table 1: Data quality objectives for ambient air quality assessment for indicative measurements in the form of given threshold percentages for various compounds. From Annex I in EC (2008).

Compound(s)	$SO_2/NO_2/CO$	Benzene	PM10/PM2.5/Lead	O_3
REU threshold in %	25	30	50	30

For those values of y for which $U_r(y)\times 100 \le \text{thr}\,\%$, where thr is the threshold in per cent for the given compound, the assessment method, i.e. low-cost sensor in our case, will comply with the EC directive (EC, 2008) in terms of indicative measurements.

The crucial aspect of using (1.1) is to evaluate or estimate the variance of the measurement error of the assessment method, i.e. our low-cost instrument. In the EC Guidance document (EC, 2010) the REU is given by the following formula

$$U_r(y_i) = \frac{2\sqrt{\frac{\text{RSS}}{n-2} - u^2(x_i) + \left(b_0 + (b_1 - 1)x_i\right)^2}}{y_i}. \tag{1.2}$$

Here n is the number of measurements, and x_i and y_i, for $i = 1,...,n$, refers to the reference and low-cost measurements, respectively. Further, b_0 and b_1 are estimated regression coefficients obtained via orthogonal regression between x and y, and RSS denotes the so-called residual sum of squares obtained after fitting the regression coefficients to the data as $\text{RSS} = \sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right)^2$.

Finally, in (1.2) $u^2(x_i)$ denotes the measurement error variance of the reference measurements. The expression under the root sign in (1.2) is thus an estimate of the variance V_i of the error of the low-cost sensor measurements y_i expressed in terms of the corresponding reference measurement x_i.

Application of orthogonal regression when comparing measurements forms a crucial part of the derivation and use of the REU formula in (1.2). It is therefore important to describe this part a little more thoroughly which is done in the next section.

2 Comparing low-cost and reference measurements using orthogonal regression

In the following, let x_i denote the reference measurements, and y_i the corresponding low-cost measurements, for a set of time points $i = 1,...,n$. We generally assume that each pair x_i, y_i can be viewed as measurements with the same sampling period, e.g. hourly or daily mean values, etc.

Associated with these measurements are the *true* underlying reference and low-cost concentrations, or measurands X_i and Y_i, related to the measurements via the following expressions

$$x_i = X_i + \varepsilon_{x_i} \quad \text{and} \quad y_i = Y_i + \varepsilon_{y_i}, \tag{2.1}$$

where ε_{x_i} and ε_{y_i} denotes the reference and low-cost measurement errors, respectively, for $i = 1,...,n$. It is important to note that the measurand values and errors are not directly observable or known, only the actual measurements x_i, y_i, for $i = 1,...,n$, are known.

The following two-variable linear regression equation is then introduced for linking the true reference concentrations with the true low-cost concentrations

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{2.2}$$

where β_0 and β_1 are the regression intercept and slope, respectively, and where u_i, for $i = 1,...,n$, represents regression residuals or errors. In the literature of methods for comparing measurements, these are also known as *equation errors*. It is assumed in orthogonal regression that $E\{u_i\} = 0$ and $\mathrm{cov}(u_i, X_i) = 0$, i.e. that the residuals have mean zero and are uncorrelated with X_i.

Using (2.1), the regression equation in (2.2) can alternatively be written

$$y_i = \beta_0 + \beta_1 x_i + v_i, \tag{2.3}$$

with residuals $v_i = u_i + \varepsilon_{y_i} - \beta_1 \varepsilon_{x_i}$. Here we still have zero mean residuals, i.e. $E\{v_i\} = 0$, but these residuals will not be independent of the x_i variables since

$$\mathrm{cov}(v_i, x_i) = \mathrm{cov}(u_i + \varepsilon_{y_i} - \beta_1 \varepsilon_{x_i}, X_i + \varepsilon_{x_i}) = -\beta_1 \sigma^2_{\varepsilon_x}. \tag{2.4}$$

Note also that since ε_{x_i} and ε_{y_i} are independent of u_i, we have

$$\mathrm{var}\, v_i = \mathrm{var}\, u_i + \beta_1^2 \sigma^2_{\varepsilon_x} + \sigma^2_{\varepsilon_y}. \tag{2.5}$$

In either (2.2) or (2.3), the regression coefficients are not known and must be estimated from the data, i.e. from the actual measurements x_i and y_i. Furthermore, the residuals are not known either. However, the variance of the residuals can be estimated using the data and the estimated regression coefficients.

The equation errors u_i or v_i are often the most important errors to consider when comparing measurements from two different instruments and are often of a much larger magnitude than the measurement errors themselves.

Due to the symmetry in the situation of comparing two instruments, one should use a regression method that respects and fully uses the symmetry between x and y. One such approach which is suggested to use in the EC Guidance document (EC, 2010) is *orthogonal regression*. The classical formulae for orthogonal regression are given in the EC Guidance document (Annex B) and are also used in the recent paper by Spinelle et al. (2015).

However, a theoretical and empirical study conducted by Dissanaike and Wang (2003) shows that the classical formulae of orthogonal regression are incorrect in that they implicitly assume a model *without* any equation errors. Consequently, as shown in their paper, the resulting estimates of the regression parameters using the classical approach will be biased and inappropriate to use in many cases, resulting in estimates of the slope coefficient which in general cannot be trusted. In fact, as stated in

their paper, by scaling the variables individually one may obtain *completely arbitrary* estimates of the slope coefficient.

Dissanaike & Wang (2003) offer a proper solution to this problem in the form of an *unbiased* estimator for the parameters of orthogonal regression. Their method is given in the form of a two-step adjusted orthogonal regression estimator, where in the first step, the variance of the residuals is estimated, and then a proper set of formulae are used in the second step to estimate the regression coefficients. They show that their method performs better than the classical estimator, and in most cases, better than ordinary least squares. Besalu et al. (2010) give an extensive theoretical and empirical study of this method in chemistry-related studies confirming these findings. The Besalu et al. (2010) paper also contains references to the use of this method in other disciplines such as in biology, economics, and physics, including aerosol science.

Appendix A contains a complete description of the two-step adjusted orthogonal regression method.

3 A derivation of the relative expanded uncertainty formula

The EC Guidance documents (EC, 2010; 2008) does not contain an explicit derivation of the REU formula (1.2). In the following, we will, therefore, attempt to derive this formula ourselves from some basic principles. The starting point of this derivation will be the difference, or error, between the low-cost sensor measurement y_i, and the true reference concentration or measurand X_i, which may be written

$$\text{error}_i = y_i - X_i = \beta_0 + \beta_1 x_i + v_i - x_i + \varepsilon_{x_i} = v_i + \varepsilon_{x_i} + \beta_0 + (\beta_1 - 1) x_i, \tag{3.1}$$

for each time point $i = 1, \ldots, n$. The aim is to obtain an expression for the variances V_i of these errors. The REU of y_i with a 95 % coverage assuming normal errors is then given by

$$U_r(y_i) = \frac{2\sqrt{V_i}}{y_i}. \tag{3.2}$$

We next derive the variance V_i. Note that (3.1) can alternatively be written

$$\text{error}_i - \varepsilon_{x_i} = v_i + \beta_0 + (\beta_1 - 1) x_i. \tag{3.3}$$

Now since the random variables error_i and ε_{x_i} are independent, we first obtain

$$\text{var}(\text{error}_i) + \sigma_{\varepsilon_x}^2 = \sigma_v^2 + \text{var}(\beta_0 + (\beta_1 - 1) x_i) + 2(\beta_1 - 1)\text{cov}(v_i, x_i). \tag{3.4}$$

Inserting (2.4) into (3.4) and rearranging we obtain

$$V_i = \text{var}(\text{error}_i) = \sigma_v^2 + \{-2\beta_1^2 + 2\beta_1 - 1\}\sigma_{\varepsilon_x}^2 + \text{var}(\beta_0 + (\beta_1 - 1) x_i). \tag{3.5}$$

Here σ_v^2 is the variance of the error term in the orthogonal regression between x_i and y_i, for $i = 1, \ldots, n$, and thus, can be estimated by using the residual sum of squares (RSS) as follows

$$\hat{\sigma}_v^2 = \frac{\text{RSS}}{n-2} = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2}{n-2}. \tag{3.6}$$

The division by $n-2$ rather than n corrects for a bias introduced by the two estimated β-coefficients in (3.6) and makes the resulting estimator of σ_v^2 unbiased.

The next term in (3.5) is $\left\{-2\beta_1^2+2\beta_1-1\right\}\sigma_{\varepsilon_x}^2$ where $\sigma_{\varepsilon_x}^2$ is the variance of the reference measurement error and assumed to be known. This term can, therefore, be estimated using

$$\left\{-2\hat{\beta}_1^2+2\hat{\beta}_1-1\right\}\sigma_{\varepsilon_x}^2. \tag{3.7}$$

The variance $\sigma_{\varepsilon_x}^2$ in (3.7) may be homoscedastic, i.e. constant and not varying with X_i, or heteroscedastic, i.e. depending on X_i, e.g. increasing with X_i. One possibility could then be to operate with the following measurement error model for this variance

$$\sigma_{\varepsilon_x}^2\left(X_i\right)=\sigma_a^2+\sigma_r^2X_i^2, \tag{3.8}$$

where σ_a^2 and σ_r^2 are known absolute and relative measurement error variances for the reference instrument. The measurement error standard deviation will then be $\approx\sigma_a$ for $X_i\approx 0$ and $\approx\sigma_r X_i$ for $X_i>>0$. For this error model, since X_i is not known, X_i needs to be replaced by e.g. x_i in the expression for the measurement error. Since

$$\begin{aligned}E\left\{\sigma_{\varepsilon_x}^2\left(x_i\right)\right\}&=E\left\{\sigma_a^2+\sigma_r^2\left(X_i+\varepsilon_{x_i}\right)^2\right\}=\sigma_a^2+\sigma_r^2E\left\{X_i^2+2X_i\varepsilon_{x_i}+\varepsilon_{x_i}^2\right\}\\&=\sigma_a^2+\sigma_r^2X_i^2+\sigma_r^2\left\{\sigma_a^2+\sigma_r^2X_i^2\right\}=\left\{1+\sigma_r^2\right\}\sigma_{\varepsilon_x}^2\left(X_i\right),\end{aligned} \tag{3.8b}$$

an unbiased estimate of this measurement error variance is then obtained using

$$\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right)=\sigma_{\varepsilon_x}^2\left(x_i\right)\Big/\left\{1+\sigma_r^2\right\}. \tag{3.8c}$$

Finally, consider the last term in (3.5), the variance of $\beta_0+\left(\beta_1-1\right)x_i$ which can be written

$$\operatorname{var}\left(\beta_0+\left(\beta_1-1\right)x_i\right)=E\left\{\left(\beta_0+\left(\beta_1-1\right)x_i\right)^2\right\}-\left\{E\left(\beta_0+\left(\beta_1-1\right)x_i\right)\right\}^2. \tag{3.9}$$

Here $E\left(\beta_0+\left(\beta_1-1\right)x_i\right)=\beta_0+\left(\beta_1-1\right)X_i$ which is zero in general only if $\beta_0=0$ and $\beta_1=1$. In this case (3.9) may be written

$$\operatorname{var}\left(\beta_0+\left(\beta_1-1\right)x_i\right)=E\left\{\left(\beta_0+\left(\beta_1-1\right)x_i\right)^2\right\}. \tag{3.10}$$

If we further assume that the variance in (3.9) or (3.10) is heteroscedastic, i.e. that it varies with x_i, then the only way to estimate this variance is to use the value $\beta_0+\left(\beta_1-1\right)x_i$ directly and square it. Thus, we obtain the following *very crude* estimate for this variance

$$\left(\hat{\beta}_0+\left(\hat{\beta}_1-1\right)x_i\right)^2. \tag{3.11}$$

Combining the expressions (3.6), (3.7) and (3.11) we obtain the following estimate of V_i

$$\hat{V}_i = \frac{\mathrm{RSS}}{n-2} + \left\{-2\hat{\beta}_1^2 + 2\hat{\beta}_1 - 1\right\}\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right) + \left(\hat{\beta}_0 + \left(\hat{\beta}_1 - 1\right)x_i\right)^2. \tag{3.12}$$

Inserting this expression into (3.2) we obtain the following REU of y_i with a 95 % coverage at the low-cost measurement level y_i

$$U_r\left(y_i\right) = \frac{2\sqrt{\frac{\mathrm{RSS}}{n-2} + \left(-2\hat{\beta}_1^2 + 2\hat{\beta}_1 - 1\right)\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right) + \left(\hat{\beta}_0 + \left(\hat{\beta}_1 - 1\right)x_i\right)^2}}{y_i}. \tag{3.13}$$

This formula is nearly the same as in the EC Guidance document (EC, 2010) and in the recent paper by Spinelle et al. (2015). If $\hat{\beta}_1 \approx 1$ or we simply assume that we perform orthogonal regression using (2.3) with the assumption that v_i and x_i are independent, i.e. $\mathrm{cov}\left(v_i, x_i\right) = 0$, we essentially obtain the same expression as in the (EC, 2010) and Spinelle et al. (2015), i.e. (1.2).

Since the assumption $\beta_0 = 0$ and $\beta_1 = 1$ generally does not hold, and thus the last term on the right-hand side of (3.9) is non-zero, the expression in (3.13) is valid generally only as an *upper bound* on the REU at the level y_i. This means that if $U_r\left(y_i\right)$ as calculated by (3.13) is less than some threshold level, then a properly calculated $U_r\left(y_i\right)$ taking into account that the second term on the right-hand side of (3.9) is non-zero will also be below this same threshold level. Thus, (3.13) enables us to state for which values of concentrations y_i the low-cost sensor complies with the Data Quality Objective (DQO) of the EC directive (EC, 2008), but it is a *conservative formula*, i.e. there might be data values failing to comply with the DQO according to this formula which nevertheless strictly speaking comply according to a more correctly calculated REU taking into account that the second term on the right-hand side of (3.9) is generally non-zero. It seems difficult, however, to come up with a better and more accurate formula.

4 The standard procedure for calculating the relative expanded uncertainty

Here we give the standard procedure or recipe for calculating the REU for a given compound based on a set of measurements x_i, y_i, for $i = 1, ..., n$, from a reference and a low-cost sensor, respectively, where n denotes the number of measurements from each instrument.

It is assumed that the ratio of variances between the low-cost and reference measurements $\lambda = \sigma_{\varepsilon_y}^2 \big/ \sigma_{\varepsilon_x}^2$ is *known*. If this is not the case, or if one is unsure about this ratio, it is recommended in the two-step adjusted orthogonal regression procedure (Dissanaike and Wang, 2003) to set $\lambda = 1$.

However, if one believes the measurement uncertainties in the low-cost instrument are higher than in the reference instrument one could try to use higher values of λ and compare the results. This recommendation also applies to the alternative procedure in Section 5.

The standard procedure now consists of the following steps:

Step 1. Perform orthogonal regression according to the two-step adjusted procedure as outlined in Appendix A. The first part of this procedure produces estimates $\tilde{\beta}_0$, $\tilde{\beta}_1$ and $\hat{\sigma}_u^2$. The second part produces final estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ based on the estimate $\hat{\sigma}_u^2$.

Step 2. Calculate the REU using

$$U_r\left(y_i\right)=\frac{2\sqrt{\frac{\mathrm{RSS}}{n-2}-\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right)+\left(\hat{\beta}_0+\left(\hat{\beta}_1-1\right)x_i\right)^2}}{y_i}, \tag{4.1}$$

where $\mathrm{RSS}=\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)$.

For each data value y_i, one then checks if $U_r\left(y_i\right)<\mathrm{thr}\,\%$, where thr is the given threshold in per cent for the current compound (see Table 1 in Section 1). Plotting the curve of such values as a function of y_i indicates for which concentration levels the low-cost sensor complies with the EC directive (EC, 2008). It is important that it complies with the directive at or above the *limit value* for the given compound.

The variance under the root sign in (4.1) is

$$\hat{V}_i=\frac{\mathrm{RSS}}{n-2}-\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right)+\left(\hat{\beta}_0+\left(\hat{\beta}_1-1\right)x_i\right)^2. \tag{4.2}$$

Thus, we see that this total variance which contributes to the REU value consist of the following three terms:

1. The regression residual error variance $\frac{\mathrm{RSS}}{n-2}$, with $\mathrm{RSS}=\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_i\right)$.
2. The reference measurement error variance $-\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right)$.
3. The error variance due to the deviation from an ideal 45-degree line, i.e. $\left(\hat{\beta}_0+\left(\hat{\beta}_1-1\right)x_i\right)^2$.

Clearly, the REU increases with increasing residual error variance as given by the first term.

Also, the REU increases when the regression line deviates more from the ideal 45-degree line corresponding to $\hat{\beta}_0=0$ and $\hat{\beta}_1=1$ as given by the third term. Thus, even without any equation or measurement errors per se, the REU will be large if the fitted regression line deviates from the ideal $y=x$ relationship. Thus, REU does not only consider measurement and equation and errors but also the systematic deviation of the fitted regression line from the ideal 45-degree line through the origin.

In (4.2) the reference measurement error variance always contributes *negatively* to the estimated total variance. Thus, a larger assumed reference measurement error variance will lead to lower REU values everything else being equal. Note, however, that the *estimated* residual error variance here *includes* both reference and low-cost measurement error variances. Thus, when the reference measurement error variance is correctly specified only the low-cost measurement error variance will remain after summing the first two terms in (4.2).

Finally, it is important to reiterate that REU might be large, irrespective of the regression residual and measurement errors if the fitted line from the orthogonal regression deviates from the ideal 45-degree line through the origin. As the results in Appendix B, Figure B.5a, shows, based on a simulation example, maximum REU will be at around 0.2 (20 %) already for $\hat{\beta}_0=2$ and $\hat{\beta}_1=0.9$ even without any residual or measurement errors.

5 An alternative procedure for calculating the relative expanded uncertainty

Here we give an alternative procedure or recipe for calculating the REU based on following more closely the model we have assumed for orthogonal regression as outlined in Section 2.

The starting point of this approach is to replace the variance of v_i, i.e. σ_v^2 in (3.5) with the more accurate formula (2.5), where $\sigma_{\varepsilon_y}^2 = \lambda\sigma_{\varepsilon_x}^2$. Thus, the variance $\hat{V}_i$ in (3.12) then becomes

$$\hat{V}_i = \hat{\sigma}_u^2 + \left(\hat{\beta}_1^2 + \lambda\right)\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right) + \left(-2\hat{\beta}_1^2 + 2\hat{\beta}_1 - 1\right)\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right) + \left(\hat{\beta}_0 + \left(\hat{\beta}_1 - 1\right)x_i\right)^2 .$$

Here σ_u^2 can be estimated using $\hat{\sigma}_u^2 = \mathrm{RSS_c}/(n-2)$ with $\mathrm{RSS_c} = \sum_{i=1}^{n}\left\{v_i^2 - \left(\hat{\beta}_1^2 + \lambda\right)\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right)\right\}$ and $v_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.

The alternative procedure then consists of the following steps:

Step 1. Perform orthogonal regression according to the two-step adjusted procedure as outlined in Appendix A. The first part of this procedure produces estimates $\tilde{\beta}_0$, $\tilde{\beta}_1$ and $\hat{\sigma}_u^2$. The second part produces final estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ based on the estimate $\hat{\sigma}_u^2$.

Step 2. Calculate the REU using the alternative formula

$$U_r\left(y_i\right) = \frac{2\sqrt{\hat{\sigma}_u^2 + \left(\hat{\beta}_1^2 + \lambda\right)\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right) + \left(-2\hat{\beta}_1^2 + 2\hat{\beta}_1 - 1\right)\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right) + \left(\hat{\beta}_0 + \left(\hat{\beta}_1 - 1\right)x_i\right)^2}}{y_i}, \quad (5.1)$$

with $\hat{\sigma}_u^2 = \mathrm{RSS_c}/(n-2)$; $\mathrm{RSS_c} = \sum_{i=1}^{n}\left\{v_i^2 - \left(\hat{\beta}_1^2 + \lambda\right)\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right)\right\}$ and $v_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$.

The formula (5.1) is more correct to use than the classical formula (4.1) under the model that we have assumed for orthogonal regression as outlined in Section 2. It becomes identical with the standard formula (4.1) when there are no measurement errors, i.e. $\hat{\sigma}_{\varepsilon_x}\left(x_i\right) = 0$.

By combining similar terms related to the measurement error, the variance under the root sign in (5.1) can be written

$$\hat{V}_i = \hat{\sigma}_u^2 + \left(\lambda - \left(\hat{\beta}_1 - 1\right)^2\right)\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right) + \left(\hat{\beta}_0 + \left(\hat{\beta}_1 - 1\right)x_i\right)^2 . \quad (5.2)$$

Thus, we see that the total variance now contributing to the REU value consist of the following three terms:

1. The equation error variance $\hat{\sigma}_u^2$.
2. The measurement error variance $\left(\lambda - \left(\hat{\beta}_1 - 1\right)^2\right)\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right)$.
3. The error variance due to the deviation from an ideal 45-degree line, i.e. $\left(\hat{\beta}_0 + \left(\hat{\beta}_1 - 1\right)x_i\right)^2$.

Clearly, the REU again increases with increasing equation error variance as given by the first term.

It is also clear that again the REU increases when the regression line deviates more from the ideal 45-degree line corresponding to $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = 1$ as given by the third term. Thus, again even without any equation or measurement errors, the REU will be large if the fitted regression line deviates from the ideal $y = x$ relationship. Thus, again REU does not only consider equation and measurement errors but also the systematic deviation of the fitted regression line from the ideal 45-degree line through the origin.

However, whether the measurement errors contribute positively to the estimated variance in (5.2) and the subsequent REU value will now depend on the sign of the expression $\lambda - \left(\hat{\beta}_1 - 1\right)^2$. Note that when $\hat{\beta}_1 \approx 1$, it will add approximately $\lambda \hat{\sigma}^2_{\varepsilon_x}\left(x_i\right)$ to the total variance, i.e. precisely the measurement error variance of y_i. Also, since usually we use $\lambda = 1$ and $1 - \left(\hat{\beta}_1 - 1\right)^2 > 0$ for all $\hat{\beta}_1 \in [0, 2]$, the second term in (5.2) will usually contribute positively to the total variance and REU.

Finally, it is again important to reiterate that REU might be large, irrespective of the equation and/or measurement errors if the fitted line from the orthogonal regression deviates from the ideal 45-degree line through the origin. As the results in Appendix B, Figure B.5b, shows, based on a simulation example, the maximum REU will be at around 0.2 (20 %) already for $\hat{\beta}_0 = 2$ and $\hat{\beta}_1 = 0.9$ even without any equation or measurement errors.

Appendix C contains a short description of an R package called `reu` implementing both the standard and the alternative procedure for calculating REU values. Also, the package contains a plotting routine for making standard REU plots based on the calculated REU values.

6 Deviations from assumptions of orthogonal regression

In this section, we discuss some possible deviations from the assumptions of orthogonal regression and how this may impact the results of orthogonal regression and the REU calculations. In the case of measurements or model error distributions not being normal, we also discuss the use of transformations to correct for this.

6.1 Incorrect specification of measurement uncertainty

As part of the orthogonal regression procedure, measurement uncertainty needs to be specified, and this needs to be done for both reference measurements x and low-cost measurements y. It is generally not possible to infer the amount of measurement uncertainty from the data, i.e. from the measurements themselves. They need to be pre-specified by the user *before* we apply the method of orthogonal regression.

Specifying measurement uncertainty for the reference measurements may be done e.g. by defining the two parameters σ_a and σ_r in the expression for the measurement error variance (3.8), or its standard deviation obtained by taking the square root of (3.8), if we believe in this error model for the reference measurements. These parameters represent respectively, the absolute measurement error standard deviation when the underlying true concentration level X_i is zero or close to zero, and the

relative standard deviation as a fraction (or per cent) of X_i for higher values of X_i. Typical values could e.g. be in the range 1-3 µgm^{-3} for σ_a, and 0.05-0.10 (5-10 %) for σ_r.

Measurement uncertainty for the low-cost sensor is specified by setting the orthogonal regression parameter λ, representing the ratio of the measurement error variance of the low-cost sensor vs. the reference instrument. As stated earlier, if we know very little or nothing about the measurement uncertainties of the low-cost sensor, it is recommended as part of the two-step adjusted orthogonal regression method of Dissanaike and Wang (2003) as outlined in Appendix A, to set this parameter to 1, i.e. to specify the same amount of uncertainty for the low-cost sensor as for the reference instrument. If we know more about the measurement error uncertainty of the low-cost instrument we might specify a different λ value, e.g. a somewhat higher value than 1, in the orthogonal regression method.

Applying orthogonal regression prevents us, however, from specifying the measurement error variance of the low-cost sensor precisely, since it always has to be defined through the use of the λ parameter, which needs to be set to a constant value for all time points. Thus, even if we manage to describe precisely the measurement uncertainty of the reference instrument, there will almost always and inevitably be some degree of approximation involved in specifying the measurement uncertainty for the low-cost sensor since it has to be made proportional to the reference uncertainty with proportionality constant λ.

One may, therefore, ask: What is the impact on orthogonal regression and the subsequent REU calculations of specifying incorrectly the measurement uncertainty of the low-cost sensor, or for that matter of the reference instrument, or both?

We consider first the case of incorrectly specifying the low-cost measurement uncertainties under the tacit assumption that the uncertainties associated with the reference measurements are specified correctly. Appendix B includes a small simulation study describing the effect of specifying incorrectly $\lambda = 1$, while in reality, its value is 2, 3, or 5. As shown in Table B.3, this seems to have a relatively small effect on the estimates of the regression coefficients β_0 and β_1, with estimates only gradually becoming further from their true values with increasing λ. We believe this to be the situation in most cases if the measurement errors are independent and with mean zero, i.e. when there are no biases in the measurements, which we assume here (see Section 6.2 for the case of bias in the measurements).

Thus, specifying incorrectly the measurement uncertainty in the y data will generally only have a modest or limited effect on the point estimates of the regression coefficients. The standard deviations of the point estimates will, however, generally increase with increasing λ, which is natural since the uncertainty increases. As seen from Table B.3 we see that the increase in parameter estimation standard deviations from this simulation study, which is based on bootstrapping, is not very dramatic.

The biggest impact, however, as seen from Table B.3, of underestimating measurement uncertainty in y is on the estimate of σ_u, the model equation standard deviation. As seen from this table, the estimate increases from about 8 for $\lambda = 1$ (which is correct) to approximately 9.5, 10.7, and 12.8 for λ equal to 2, 3, and 5, respectively. Thus, the impact on orthogonal regression of specifying a too small uncertainty in y is to inflate the estimate of the equation error standard deviation σ_u. This is evident also from the estimation equation of σ_u as given in equation (A.6) in Appendix A, which can be written

$$\hat{\sigma}_u^2 = \frac{\mathrm{RSS_c}}{n-2}; \quad \mathrm{RSS_c} = \sum_{i=1}^{n}\left\{v_i^2 - \tilde{\beta}_1^2\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right) - \hat{\sigma}_{\varepsilon_y}^2\left(x_i\right)\right\}; \quad v_i = y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i \tag{6.1}$$

Thus, as long as the estimated parameters $\tilde{\beta}_0$ and $\tilde{\beta}_1$ are relatively unaffected by the incorrectly specified measurement uncertainty in y, which they generally are, the impact of too small $\hat{\sigma}_{\varepsilon_y}^2\left(x_i\right)$ values for $i = 1,...,n$ will be a too high value of $\hat{\sigma}_u$; and vice versa, if the $\hat{\sigma}_{\varepsilon_y}^2\left(x_i\right)$ values are too high, the effect will be a too low value of $\hat{\sigma}_u$.

From the expression (6.1) we also see that incorrect specification of the measurement uncertainties of the reference instrument $\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right)$ will have a similar effect on $\hat{\sigma}_u$, i.e. that too small $\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right)$ values will lead to a too high value of $\hat{\sigma}_u$; and vice versa, if $\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right)$ values are too high, this leads to a too low value of $\hat{\sigma}_u$.

Thus, to summarise, if we *underestimate* the measurement uncertainty of either instrument, we will *overestimate* the model equation uncertainty; and if we *overestimate* the measurement uncertainty in either instrument, we will *underestimate* the model equation uncertainty.

The latter case has an important practical implication: If we obtain an estimate $\hat{\sigma}_u = 0$, it indicates we operate with too large measurement uncertainties in either the reference instrument, or in the low-cost sensor, or both.

The impact of incorrectly specifying measurement uncertainty on the standard REU values is easily seen from the expression for standard REU as given by equation (4.1).

Since the standard REU only depends on $\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right)$, and not on $\hat{\sigma}_{\varepsilon_y}^2\left(x_i\right)$, there will be no impact on the standard REU values of incorrectly specifying measurement uncertainties for the low-cost instrument as long as the estimates of the regression coefficients are fairly unaffected.

Underestimation of measurement uncertainties in the reference instrument, however, will lead to an *overestimation* of the standard REU values; and *overestimation* of measurement uncertainties in the reference instrument leads to an *underestimation* of the standard REU values.

Again, this has an important practical implication: Since REU values all should be positive, obtaining standard REU values equal to zero indicates we operate with too large measurement uncertainties in the reference instrument.

6.2 Measurement bias

The model for orthogonal regression as outlined in Section 2 assumes unbiased measurement errors for both x and y. We will here briefly consider the consequences of performing orthogonal regression when there are biases in one or both measurements. Assume therefore that the measurement equations (2.1) are replaced by

$$x_i = X_i + \mu_i + \varepsilon_{x_i} \quad \text{and} \quad y_i = Y_i + \nu_i + \varepsilon_{y_i}, \tag{6.2}$$

where μ_i and ν_i are the biases of the x_i and y_i measurements, respectively, at each time point i, for $i = 1,...,n$. We may view (6.2) as still representing unbiased measurements but now for a pair of

alternative or perturbed underlying true concentrations or measurands $X_i' = X_i + \mu_i$ and $Y_i' = Y_i + \nu_i$ for $i = 1,...,n$.

If the biases are constant with time, i.e. $\mu_i = \mu$ and $\nu_i = \nu$, then the model equation (2.2) can be written

$$Y_i' - \nu = \beta_0 + \beta_1\left(X_i' - \mu\right) + u_i \quad \Leftrightarrow \quad Y_i' = \beta_0' + \beta_1 X_i' + u_i, \tag{6.3}$$

with $\beta_0' = \beta_0 + \nu - \beta_1\mu$, and where β_1 and u_i are unaltered. Thus, a constant bias in one or both measurements will lead to a perturbed orthogonal regression equation where only the constant term is affected. For example, if β_0 in the relationship between the measurands is zero or close to zero, the perturbed β_0' due to bias might further from zero, unless $\nu - \beta_1\mu$ is zero or close to zero.

In the more general case of biases that varies with time, estimates of all three quantities β_0, β_1 and σ_u^2 will generally be influenced by such biases. Thus, if we believe that there might be some biases in the measurements, e.g. perhaps most likely in the low-cost instrument, we should interpret the estimated parameters $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}_u^2$ not only as parameters in the linear regression relationship between the measurands associated with the instruments but in a linear regression taking into account the *combined effect of different measurands and biases in the instruments*. Thus, a $\hat{\beta}_0$ differing from zero and a $\hat{\beta}_1$ differing from 1 need not necessarily be due to a discrepancy between the measurands in the two instruments but might likely be also due to biases in one or both of the instruments, or in their combined effect.

It follows from this that biases in the measurements might easily affect the REU values and affect the conclusion as to whether two instruments are equivalent according to the DQO. This is how it should be: Biases in the low-cost sensor measurements e.g., should be reflected in increased REU values.

6.3 Non-normal measurements or model errors

Often in air quality analysis, measurement or model errors will not be normal but have distributions that are somewhat non-symmetric or skewed, typically with a longer tail to the right. In orthogonal regression we assume these errors to be normally distributed. This is also the assumption in the REU calculations and is the reason for the factor 2 in the REU expression (e.g. in (4.1) or (5.1)).

Thus, if the distribution of errors is not normal, the numerator in the REU expression may no longer be associated with a 95% coverage or confidence interval for the low-cost measured value y_i. The REU values might then correspond to a smaller coverage or confidence interval than 95% which will invalidate the use of these values to check for compliance with the DQO or testing for equivalence of instruments.

To make such errors more symmetric and normal-like, we may use transformations of the variables. One suitable transformation to use in this regard is the Box-Cox power transformation, which reads

$$z^{(\gamma)} = \begin{cases} \dfrac{z^\gamma - 1}{\gamma} + \gamma & 0 < \gamma \le 1 \\ \log z & \gamma = 0 \end{cases}, \tag{6.4}$$

where z is the variable to be transformed and γ is the Box-Cox power transformation parameter. For $\gamma = 1$ we have no transformation, i.e. $z^{(1)} = z$, and for $\gamma = 0$ we have the log-transformation, i.e. $z^{(0)} = \log z$, obtained mathematically in the limit as $\gamma \to 0+$ using L'Hôpitals rule.

If we decide to transform the variables of orthogonal regression using this transformation we obtain the following transformed measurement and model equations

$$x_i^{(\gamma)} = X_i^{(\gamma)} + \varepsilon_{x_i}; \quad y_i^{(\gamma)} = Y_i^{(\gamma)} + \varepsilon_{y_i}; \quad Y_i^{(\gamma)} = \beta_0 + \beta_1 X_i^{(\gamma)} + u_i, \tag{6.5}$$

corresponding to (2.1) and (2.2) involving the transformed measurements and measurands. For some suitable $\gamma \in [0,1]$, e.g. $\gamma = 0$ using the log-transformation, the errors involved in (6.5) will become more symmetric and normally distributed. Thus, we may apply orthogonal regression and REU calculations using the transformed measurements $x_i^{(\gamma)}$ and $y_i^{(\gamma)}$ rather than the original ones.

We will then check for compliance of the DQO in the transformed space of concentrations. It should, however, not cause problems making the same statements of compliance or equivalence of the two instruments in the transformed space as in the original space if the transformation we apply is monotone and one-to-one, which is the case using (6.4).

Note that the transformation (6.4) may produce negative values $z^{(\gamma)}$, but only if $z < 1$. Such negatively transformed values are no problem in the orthogonal regression procedure per se, but to avoid complications with the REU calculations where we divide by the transformed quantity $y_i^{(\gamma)}$, we may wish to retain only the data pairs x_i, y_i for which the low-cost measurement y_i is > 1 to avoid division by zero or negative values in these calculations.

One of the benefits of using the transformation (6.4) is that it often tends to stabilize the variance of the error terms, i.e. making the error variances more homoscedastic, i.e. constant and not varying with the level of the measurements or measurands. Thus, we may assume an error model like (3.8) for the transformed reference measurements using only $\sigma_a > 0$ and setting $\sigma_r = 0$.

A central question in connection with the Box-Cox transformation (6.4) is how to choose the power parameter γ. Here we may use the residuals v_i from (2.3), i.e.

$$v_i = y_i^{(\gamma)} - \beta_0 - \beta_1 x_i^{(\gamma)} = u_i + \varepsilon_{y_i} - \beta_1 \varepsilon_{x_i}. \tag{6.6}$$

If γ is correctly set then u_i, ε_{y_i} and ε_{x_i} should all be approximately normally distributed with constant variances. Thus, v_i in (6.6) should also be approximately normally distributed with a constant variance. By plotting the distribution of the residuals $\hat{v}_i = y_i^{(\gamma)} - \hat{\beta}_0 - \hat{\beta}_1 x_i^{(\gamma)}$ from the orthogonal regression procedure, we may visually inspect to see if these residuals look symmetric and normal-like or skewed. In the latter case, if the distribution is skewed with a longer tail to the right, γ is too large, and vice versa, if the distribution has a longer tail to the left, γ is too small. A search for a best or optimal γ may then be performed iteratively, recalculating the orthogonal regression estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ and calculating new residuals for each new value of γ.

7 A real case example

Here we apply orthogonal regression and REU calculations to a real case example taken from the EU H2020 project HackAIR (www.hackair.eu) recently published in Liu et al. (2019). This example involves comparing three low-cost Nova SDS011 PM sensors with a reference TEOM (Tapered Element Oscillating Microbalance) instrument for measuring hourly-average PM2.5 concentrations at the road-side station Kirkeveien in Oslo, Norway for the period 11 Dec 2017 to 31 Mar 2018.

In Liu et al. (2019), orthogonal regression using a standard approach (i.e. not according to the method of Dissanaike and Wang (2003) as recommended here) was performed followed by applying standard REU calculations using the expression (4.1). The results for the three low-cost sensors named ID-1, ID-2, and ID-3 are shown in Figure 13 in Liu et al. (2019) as plots of hourly REU values against hourly reference measurements with the TEOM instrument. This figure is reproduced below as Figure 7.1.

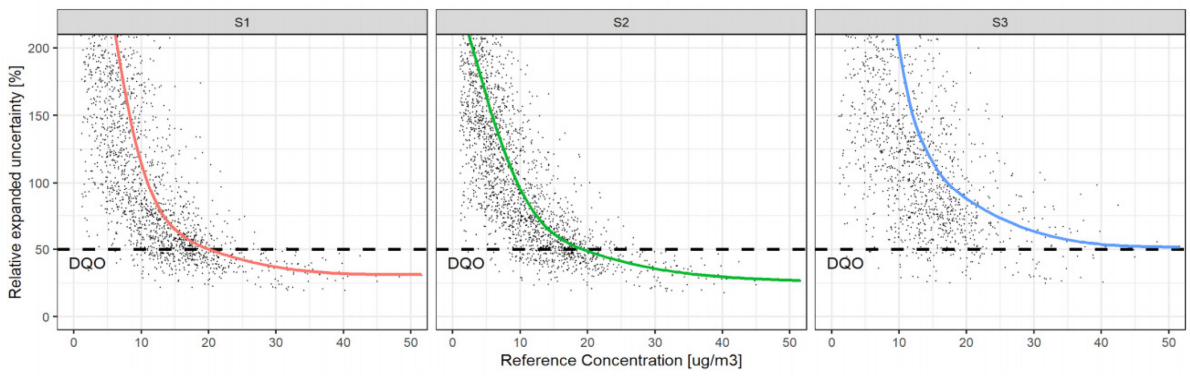


Figure 7.1:A plot of hourly REU values from Liu et al. (2019) for the low-cost sensors ID-1 (left), ID-2 (middle) and ID-3 (right).

In this plot, the hourly values are plotted as the small dots while the coloured lines represent the fit of a generalized additive model (GAM) to the data. The dashed black line represents the data quality objective (DQO) for indicative measurements for PM2.5 which is 0.5 (50 %). As can be seen from these plots and based on the coloured lines, sensors ID-1 and ID-2 may be considered as equivalent with the TEOM for concentrations above approximately 20 µgm^{-3}, while sensor ID-3 fails to be equivalent regardless of the level of concentration.

Table 7.1 shows the results of applying the two-step adjusted orthogonal regression method of Dissanaike and Wang (2003) as outlined in Appendix A to these data.

Table 7.1: Estimated parameters with standard deviation (sd) for the two-step adjusted orthogonal regression method.

Sensor	n	$\hat{\beta}_0\,(\mathrm{sd})$	$\hat{\beta}_1\,(\mathrm{sd})$	$\hat{\sigma}_u\,(\mathrm{sd})$	$\hat{\sigma}_v\,(\mathrm{sd})$
ID-1	2458	0.664 (0.138)	0.924 (0.017)	3.709 (0.130)	4.148 (0.105)
ID-2	2413	2.696 (0.144)	0.845 (0.017)	3.703 (0.141)	4.108 (0.114)
ID-3	2390	0.727 (0.170)	0.723 (0.021)	4.478 (0.175)	4.663 (0.138)

In these calculations we have assumed a measurement error model for the TEOM instrument using (3.8) with $\sigma_a = 1$ µgm^{-3} as absolute measurement error standard deviation at 0 µgm^{-3} levels; $\sigma_r = 0.1$ i.e. a 10 % relative measurement error standard deviation; and $\lambda = 1$, i.e. the same measurement error standard deviation for the low-cost sensors as for the TEOM instrument.

In Table 7.1, the standard deviations (sd) are based on bootstrapping using $B = 1000$ resamples of the original data.

The parameter estimate $\hat{\sigma}_v$ is the table is calculated by

$$\hat{\sigma}_v^2 = \frac{\mathrm{RSS}}{n-2}; \quad \mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

i.e. using the standard REU formula (4.1).

In Figs. 7.2a – 7.4b below we give plots of the relative expanded uncertainty (REU) for ID-1, ID-2, and ID-3 as calculated by the standard formula (4.1) and by the alternative and more correct formula (5.1).

In these plots only REU values less than 200 % for $x_i \geq 1$ µgm^{-3} are shown. The horizontal black dashed line corresponds to the DQO for indicative methods for PM2.5 (50 %), while the red solid line shows the corresponding line for fixed measurements (25 %). The latter line is perhaps not so relevant here since the low-cost sensor measurements are thought to be used only as indicative.

As can be seen, the plots based on the standard formula, i.e. Figs. 7.2a, 7.3a, and 7.4a are quite like the plots in Figure 7.1. Again, the REU value is below the critical 50 % when concentration levels are above 20 – 25 µgm^{-3} for both ID-1 and ID-2, while for ID-3 the REU values stay above 50 % for all levels and thus, this sensor is not in compliance with the DQO.

If we compare with the plots based on the alternative formula, i.e. Figs. 7.2b, 7.3b, and 7.4b, we see that the alternative formula gives REU values which are generally somewhat higher than the standard ones, thus the alternative method is slightly more conservative in its assessment. Still according to these plots, however, ID-1 and ID-2 are mostly in compliance with the DQO for levels above ca. 25 µgm^{-3}, while ID-3 fails to be compliant.

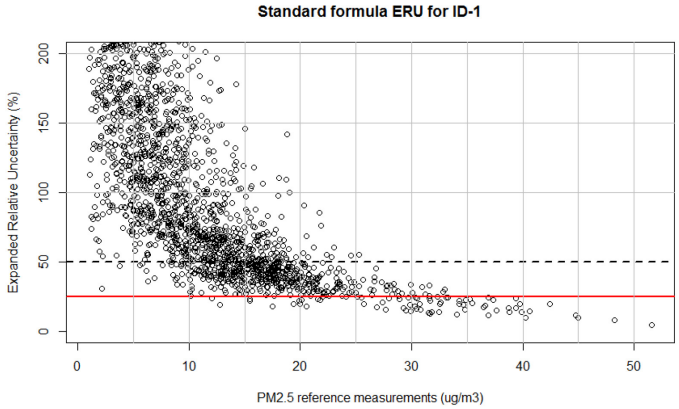


Figure 7.2a: The plot of standard formula REU for ID-1.

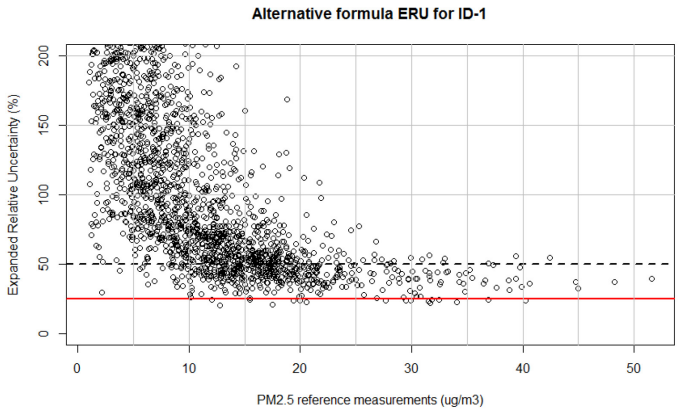


Figure 7.2b: The plot of alternative formula REU for ID-1.

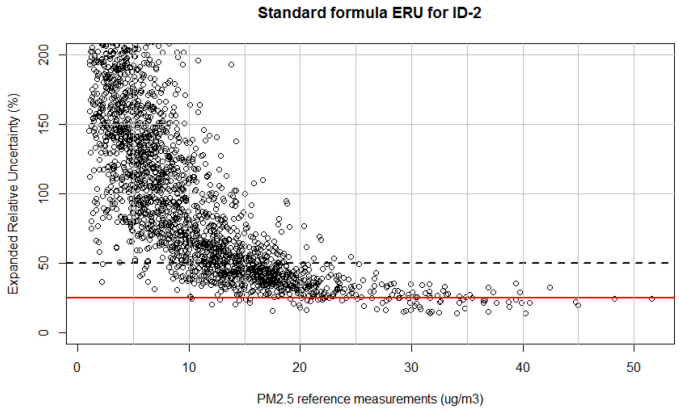


Figure 7.3a: The plot of standard formula REU for ID-2.

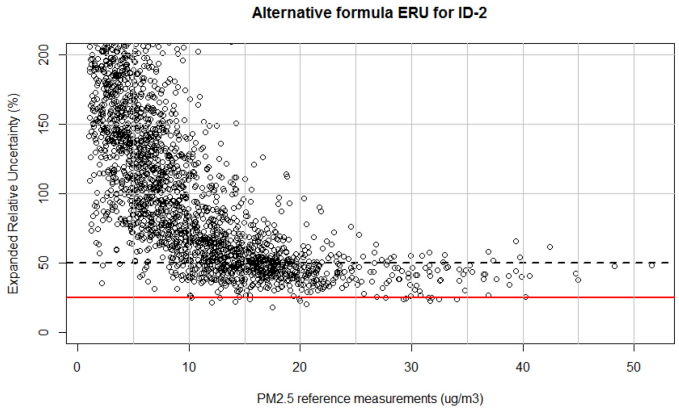


Figure 7.3b: The plot of alternative formula REU for ID-2.

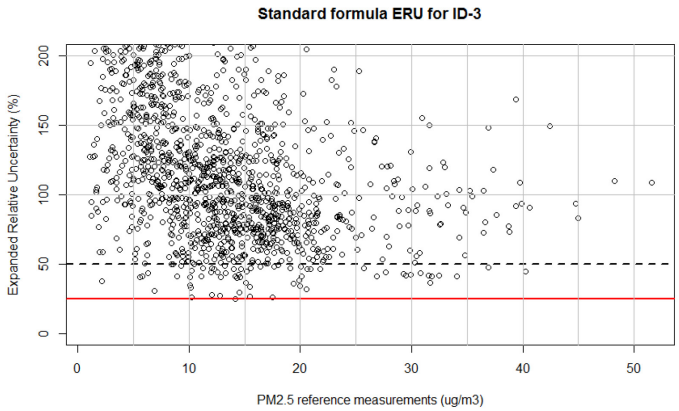


Figure 7.4a: The plot of standard formula REU for ID-3.

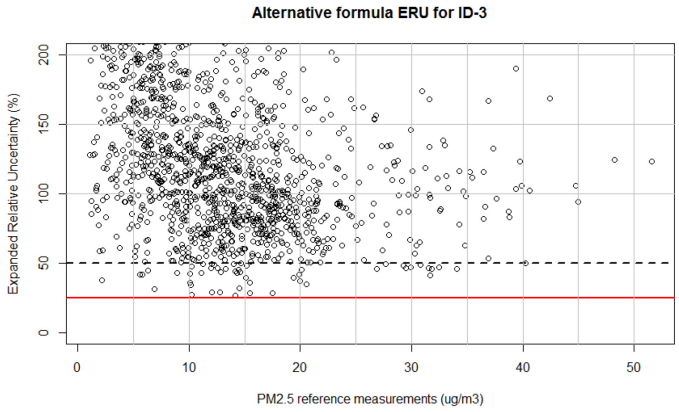


Figure 7.4b: The plot of alternative formula REU for ID-3.

8 Concluding remarks

This project aimed to gain better insight and an improved understanding of the relative expanded uncertainty (REU) formula as described in the EC Guidance document (EC, 2010) and used in the recent paper by Spinelle et al. (2015). This formula is used to define relative uncertainties of low-cost sensors for checking whether the Data Quality Objective (DQO) of the European Air Quality Directive (2008/50/EC) (EC, 2008) for indicative methods is reached for these sensors as compared with reference measurements.

As shown in this report, the standard REU formula is correct under the usual assumptions of orthogonal regression but needs an additional assumption of a statistically expected zero difference or bias between the measurands of the instruments at each time point, to represent a proper two times relative standard deviation value. The latter assumption is in practice equivalent to assuming an ideal $Y = X + u$ relationship between the low-cost and reference instrument measurands. If this assumption is not correct, a value calculated with the standard REU formula may still be used but must then be interpreted as an *upper bound* of the REU at the indicated low-cost measurement levels.

9 References

Besalu, E., de Julian-Ortiz, J. V., Pogliani, L. (2009) Ordinary and orthogonal regressions in QSAR/QSPR and chemistry-related studies. *MATCH Commun. Math. Comput. Chemistry*, 63, 573-583

Dissanaike, G. and Wang, S. (2003) A critical examination of orthogonal regression. Social Science Research Network, Electronic Paper No. 407560, 1-39.

EC (2008), Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe. 2008 O.J. L 152/1.

EC (2010), Guide to the demonstration of equivalence of ambient air monitoring methods. Report by an EC Working Group on Guidance.

Liu, H-Y, Schneider, P., Haugen, R., Vogt, M. (2019) Performance Assessment of a Low-Cost PM2.5 Sensor for a near Four-Month Period in Oslo, Norway. *Atmosphere* 2019, 10, 41.

Spinelle, L., Gerboles, M., Villani, M.G., Aleixandre, M., Bonavitacola, F. (2015) Field calibration of a cluster of low-cost available sensors for air quality monitoring. Part A: Ozone and nitrogen dioxide. *Sensors and Actuators* B 215, pp. 249-257.

Appendix A

The two-step adjusted orthogonal regression method

Here we give the formulae for the two-step adjusted orthogonal regression method as presented in Dissanaike and Wang (2003). According to their paper, they recommend this method as a practical approach for performing orthogonal regression between two variables x and y when there are measurement errors in both, and also errors associated with the linear regression equation relating these variables, but little or no prior information about the variances of the errors involved.

Assume in the following that the measurements are denoted by x_i and y_i, for $i = 1,...,n$, where n is the number of measurements. The linear regression model with measurement and equation errors is then given by the following set of equations

$$x_i = X_i + \varepsilon_{x_i}; \quad y_i = Y_i + \varepsilon_{y_i}; \quad Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{A.1}$$

where X_i and Y_i represents measurands, i.e. underlying true concentration values that would have been obtained by the instruments if there were no measurement errors ε_{x_i} and ε_{y_i}, for $i = 1,...,n$. In (A.1), the parameters β_0 and β_1 represents the intercept and slope of the regression equation and u_i the corresponding regression equation errors.

The basic assumptions behind orthogonal regression are then

A1. ε_{x_i}, ε_{y_i} and u_i are normally distributed variables for $i = 1,...,n$.

A2. ε_{x_i}, ε_{y_i} and u_i have mean value zero, i.e. $E\varepsilon_{x_i} = E\varepsilon_{y_i} = Eu_i = 0$ for $i = 1,...,n$.

A3. ε_{x_i}, ε_{y_i} and u_i are independent of X_i, Y_i and X_i, respectively, i.e. $\mathrm{cov}\left(\varepsilon_{x_i}, X_i\right) = \mathrm{cov}\left(\varepsilon_{y_i}, Y_i\right) = \mathrm{cov}\left(u_i, X_i\right) = 0$ for $i = 1,...,n$.

Let the variances of the measurement errors be $\sigma^2_{\varepsilon_x}$ and $\sigma^2_{\varepsilon_y}$, respectively, and let their variance ratio be

$$\lambda = \frac{\sigma^2_{\varepsilon_y}}{\sigma^2_{\varepsilon_x}}. \tag{A.2}$$

If one or both measurement error variances are unknown, it is suggested to simply set $\lambda = 1$.

In the following, we will assume λ has been set, either to the actual ratio of the measurement error variances if these are known, or to the value 1 if one or both of them are unknown.

In the first step of the method one estimates the intercept and slope β_0 and β_1 of the regression equation using the following standard orthogonal regression formulae (Dissanaike and Wang, 2003, Eq. 6, p. 5)

$$\tilde{\beta}_1 = \frac{s_y^2 - \lambda s_x^2 + \sqrt{\left(s_y^2 - s_x^2\right)^2 + 4\lambda s_{xy}^2}}{2s_{xy}} \tag{A.3}$$

and

$$\tilde{\beta}_0 = \overline{y} - \tilde{\beta}_1 \overline{x}\,, \tag{A.4}$$

with s_{xy} the sample covariance between x and y, and with $\overline{x}$ and $\overline{y}$ sample means and s_x^2 and s_y^2 sample variances of x and y, respectively. These quantities are calculated from the measurements as follows

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i; \quad s_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2; \quad s_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2; \quad s_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

(A.5)

An estimator for the variance of the model equation error σ_u^2 is then given by

$$\hat{\sigma}_u^2 = \frac{\mathrm{RSS_c}}{n-2}; \quad \mathrm{RSS_c} = \sum_{i=1}^{n}\left\{v_i^2 - \left(\tilde{\beta}_1^2 + \lambda\right)\hat{\sigma}_{\varepsilon_x}^2\left(x_i\right)\right\}; \quad v_i = y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i\,. \tag{A.6}$$

In the second step of their method, one again estimates the regression parameters but now using the following adjusted orthogonal regression formulae (Dissanaike and Wang (2003, Eq. 8, p. 8)

$$\hat{\beta}_1 = \frac{s_y^2 - \lambda s_x^2 - \hat{\sigma}_u^2 + \sqrt{\left(s_y^2 - \lambda s_x^2 - \hat{\sigma}_u^2\right)^2 + 4\lambda s_{xy}^2}}{2 s_{xy}} \tag{A.7}$$

and

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}\,. \tag{A.8}$$

As shown in their paper, this results in unbiased estimates of the true regression parameters, i.e. we have

$$E\left\{\hat{\beta}_0\right\} = \beta_0; \quad E\left\{\hat{\beta}_1\right\} = \beta_1\,, \tag{A.9}$$

where $E\{\cdot\}$ denotes statistical expectation. This contrasts with the standard orthogonal regression formulae (A.3) and (A.4) which do not consider properly the equation errors.

As shown by Dissanaike and Wang (2003) using the standard approach means that a unit regression slope can always be obtained by simply rescaling one of the variables, thus rendering the results virtually useless. Only the above-adjusted two-step orthogonal regression formulae assure unbiased estimates of these parameters and thus a proper use of the orthogonal regression methodology.

Appendix B

Some simulation results

Here we give some results using the two-step adjusted orthogonal regression method to calculate relative expanded uncertainties (REUs) based on sets of simulated data.

As a starting point for these calculations, true reference concentrations or measurands X_i, for $i = 1,...,n$, were defined realistically for a whole year, i.e. for $n = 8760$ hourly values by using actual measurement data.

We then simulated reference measured values $x_i = X_i + \varepsilon_{x_i}$, $i = 1,...,n$, with $\varepsilon_{x_i} = N\left(0, \sigma^2_{\varepsilon_x}\right)$ and $\sigma^2_{\varepsilon_x} = \sigma^2_a + \sigma^2_r X^2_i$, with σ_a and σ_r as given below. Next, true low-cost sensor concentrations or measurands Y_i were simulated using $Y_i = 5.0 + 0.7X_i + u_i$, with $u_i = N\left(0, \sigma^2_u\right)$, and σ_u as given below. Finally, low-cost measured values y_i were simulated using $y_i = Y_i + \varepsilon_{y_i}$ with $\varepsilon_{y_i} = N\left(0, \sigma^2_{\varepsilon_y}\right)$ and $\sigma^2_{\varepsilon_y} = \lambda\sigma^2_{\varepsilon_x}$ with $\lambda = 1$.

The following cases were considered:

Case 1: No measurement errors and no equation errors.

In this case, we set $\sigma_a = \sigma_r = 0$ and $\sigma_u = 0$. Table B.1 row 1 gives the results in this case.

Case 2: Measurement errors but no equation errors.

In this case, we set $\sigma_a = 3$, $\sigma_r = 0.1$ and $\sigma_u = 0$. Table B.1 row 2 gives the results in this case.

Case 3: No measurement errors but with equation errors.

In this case, we set $\sigma_a = \sigma_r = 0$ and $\sigma_u = 8$. Table B.1 row 3 gives the results in this case.

Case 4: Both measurement and equation errors.

In this case, we set $\sigma_a = 3$, $\sigma_r = 0.1$ and $\sigma_u = 8$. Table B.1 row 4 gives the results in this case.

Table B.1: Estimated parameters with standard deviation (sd) for the two-step adjusted orthogonal regression method based on $n = 8760$ simulated data for each case.

Case	$\hat{\beta}_0\,(\mathrm{sd})$	$\hat{\beta}_1\,(\mathrm{sd})$	$\hat{\sigma}_u\,(\mathrm{sd})$	$\hat{\sigma}_v\,(\mathrm{sd})$
1	5.000 (0.000)	0.700 (0.000)	0.000 (0.000)	0.000 (0.000)
2	5.047 (0.126)	0.696 (0.007)	0.000 (0.239)	5.677 (0.121)
3	5.017 (0.111)	0.701 (0.003)	8.074 (0.061)	8.074 (0.061)
4	5.249 (0.186)	0.689 (0.009)	8.108 (0.150)	9.898 (0.129)

In this table, the standard deviations (sd) are based on bootstrapping using $B = 1000$ resamples of the original simulated data. The parameter estimate $\hat{\sigma}_v$ is defined by

$$\hat{\sigma}^2_v = \frac{\mathrm{RSS}}{n-2}; \quad \mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

i.e. using the standard REU formula (4.1).

In Figs. B.1a – B.4b we give plots of the relative expanded uncertainty (REU) in % for these cases as calculated by the standard formula (4.1) and by the alternative and more correct formula (5.1). In these plots only REU values for $x_i \geq 1$ are shown. Note that the standard and alternative formulae are identical for Cases 1 and 3 thus producing identical plots.

As can be seen from these plots the simulated y data instrument is not in compliance with the DQO for being equivalent with the x data instrument for fixed methods or indicative methods for the higher values of reference concentrations for any of these cases as indicated by the solid red lines and dashed black lines, respectively. The reason for this is that the coefficients $\beta_0 = 5$ and $\beta_1 = 0.7$ are too far from their ideal values of 0 and 1, respectively.

To see if using e.g. $\beta_0 = 2$ and $\beta_1 = 0.9$ gives results in compliance with the DQO, we repeated the above simulations using these as parameters of the regression equation instead. The results are given in Table B.2 and Figs. B.5a – B.8b.

Table B.2: Estimated parameters with standard deviation (sd) for the two-step adjusted orthogonal regression method based on $n = 8760$ simulated data using $\beta_0 = 2$ and $\beta_1 = 0.9$.

Case	$\hat{\beta}_0(\mathrm{sd})$	$\hat{\beta}_1(\mathrm{sd})$	$\hat{\sigma}_u(\mathrm{sd})$	$\hat{\sigma}_v(\mathrm{sd})$
1	2.000 (0.000)	0.900 (0.000)	0.000 (0.000)	0.000 (0.000)
2	2.134 (0.140)	0.895 (0.008)	0.000 (0.500)	6.358 (0.151)
3	2.019 (0.111)	0.901 (0.003)	8.080 (0.062)	8.081 (0.062)
4	2.252 (0.198)	0.889 (0.010)	8.120 (0.169)	10.255 (0.141)

Now we see that we are following the DQO for higher values of reference concentrations using the standard REU formula, but not for cases 2 and 4 according to the alternative REU formula.

Next, we check to see what happens if the ratio λ between measurement error variances of y vs. x is specified incorrectly, e.g. if the measurement errors in the low-cost sensor y are much higher than those in the reference instrument x. In Table B.3 we give the results of applying both measurement and equation errors as in Table B.2, Case 4, but using simulated y observations for λ set to 2, 3, and 5, respectively.

Table B.3: Estimated parameters with standard deviation (sd) for the two-step adjusted orthogonal regression method for Case 4 ($n = 8760$) using $\beta_0 = 2$, $\beta_1 = 0.9$ and λ = 2, 3, and 5.

λ	$\hat{\beta}_0(\mathrm{sd})$	$\hat{\beta}_1(\mathrm{sd})$	$\hat{\sigma}_u(\mathrm{sd})$	$\hat{\sigma}_v(\mathrm{sd})$
2	2.316 (0.235)	0.885 (0.012)	9.481 (0.226)	11.311 (0.193)
3	2.372 (0.267)	0.882 (0.014)	10.689 (0.275)	12.279 (0.238)
5	2.481 (0.322)	0.877 (0.017)	12.813 (0.359)	14.020 (0.312)

As can be seen from this table, the three first parameter values are now somewhat further away from their true values 2, 0.9, and 8, respectively, thus the standard REU values will be slightly larger for these values of λ than for the simulations we did with $\lambda = 1$.

Finally, we check to see what happens if σ_r defining the measurement error variances of x is specified incorrectly, i.e. if the measurement errors in the reference instrument are smaller or larger than the assumed level based on the value $\sigma_r = 0.1$.

In Table B.4 we give the results of applying both measurement and equation errors as in Table B.2, Case 4, but using simulated x and y observations based on σ_r set to 0.05, 0.10, and 0.15, respectively (keeping the value $\sigma_a = 3$).

Table B.4: Estimated parameters with standard deviation (sd) for the two-step adjusted orthogonal regression method for Case 4 ($n = 8760$) using $\beta_0 = 2$, $\beta_1 = 0.9$ and σ_r = 0.05, 0.1, and 0.15.

σ_r	$\hat{\beta}_0\,(\mathrm{sd})$	$\hat{\beta}_1\,(\mathrm{sd})$	$\hat{\sigma}_u\,(\mathrm{sd})$	$\hat{\sigma}_v\,(\mathrm{sd})$
0.05	1.768 (0.135)	0.916 (0.005)	6.798 (0.121)	9.316 (0.072)
0.10	1.976 (0.177)	0.907 (0.009)	8.070 (0.133)	10.238 (0.112)
0.15	2.338 (0.229)	0.890 (0.012)	9.811 (0.206)	11.580 (0.186)

The results in row 2 are not the same as in Table B.2 row 4 since a different random number seed was used in these simulations. As can be seen from Table B.4, the first three parameter values are again somewhat further away from their true values 2, 0.9, and 8, respectively, implying that the standard REU values will again be affected, i.e. slightly increased for $\sigma_r = 0.05$ and slightly decreased for $\sigma_r = 0.15$ as compared with $\sigma_r = 0.1$.

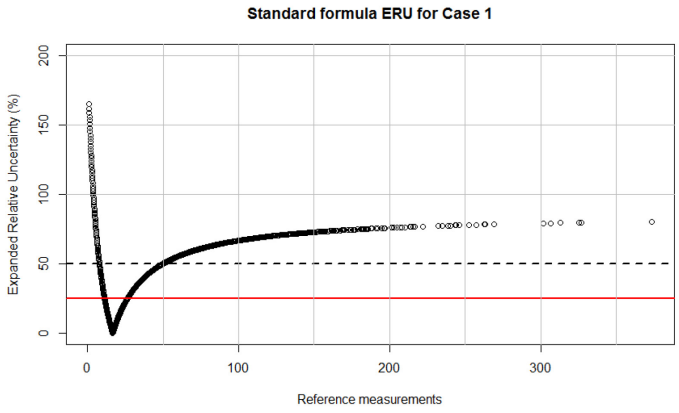


Figure B.1a: The plot of standard formula REU in % for Case 1 with $\beta_0 = 5$ and $\beta_1 = 0.7$.

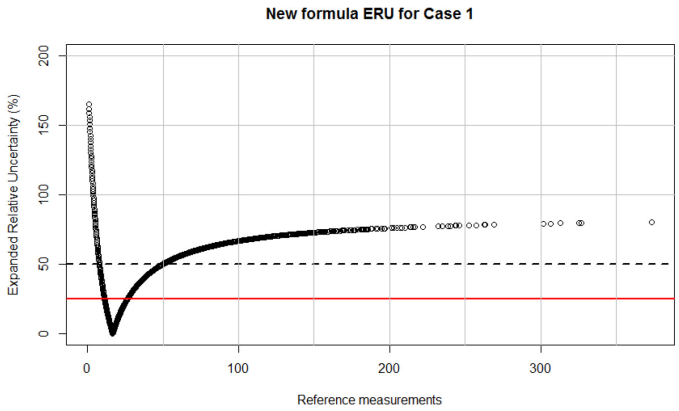


Figure B.1b: *The plot of alternative formula REU in % for Case 1 with* $\beta_0 = 5$ *and* $\beta_1 = 0.7$ *.*

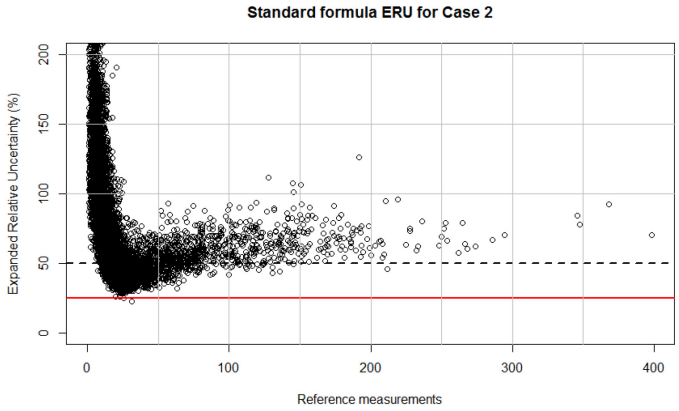


Figure B.2a: *The plot of standard formula REU in % for Case 2 with* $\beta_0 = 5$ *and* $\beta_1 = 0.7$ *.*

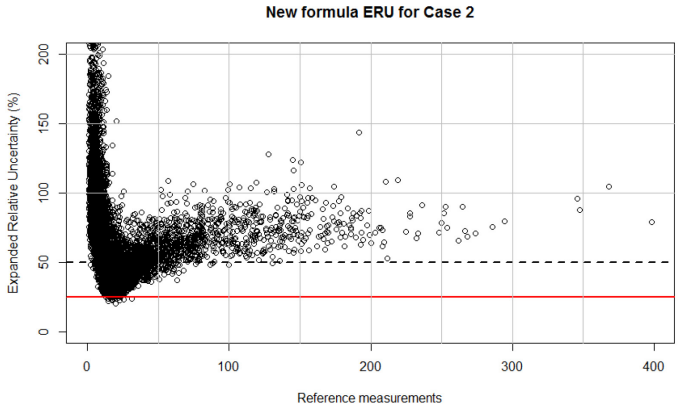


Figure B.2b: *The plot of alternative formula REU in % for Case 2 with* $\beta_0 = 5$ *and* $\beta_1 = 0.7$ *.*

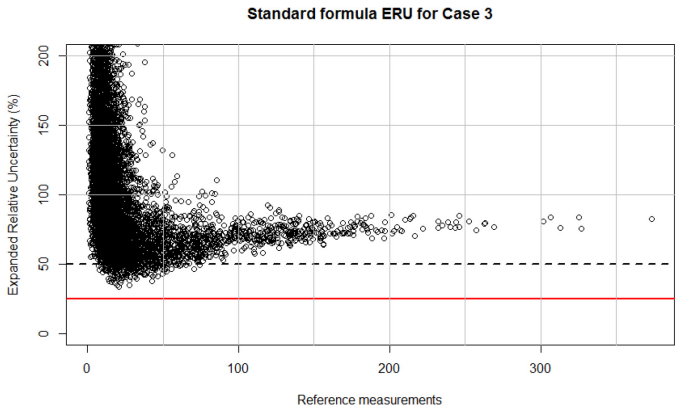


Figure B.3a: *The plot of standard formula REU in % for Case 3 with* $\beta_0 = 5$ *and* $\beta_1 = 0.7$ *.*

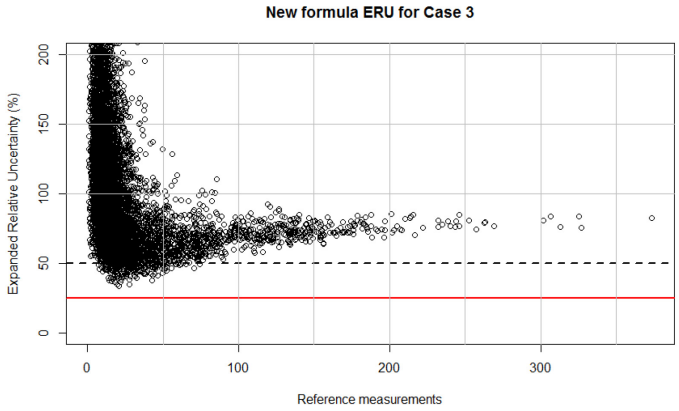


Figure B.3b: *The plot of alternative formula REU in % for Case 3 with* $\beta_0 = 5$ *and* $\beta_1 = 0.7$ *.*

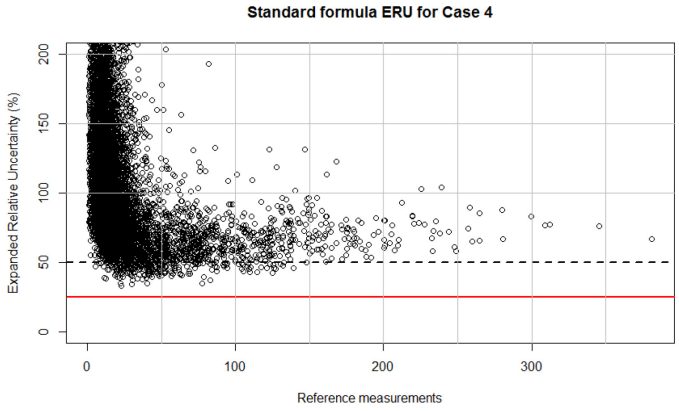


Figure B.4a: *The plot of standard formula REU in % for Case 4 with* $\beta_0 = 5$ *and* $\beta_1 = 0.7$ *.*

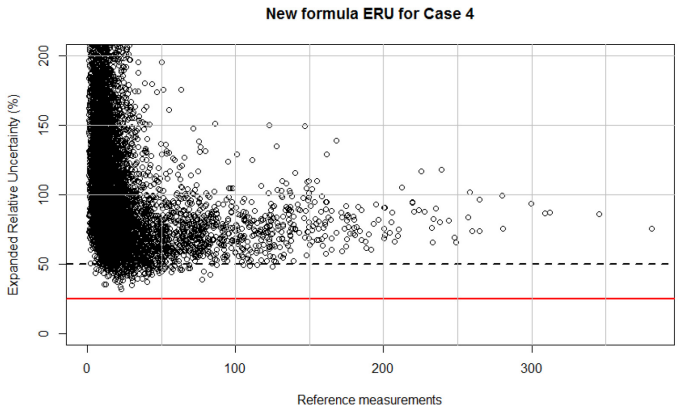


Figure B.4b: *The plot of alternative formula REU in % for Case 4 with* $\beta_0 = 5$ *and* $\beta_1 = 0.7$.

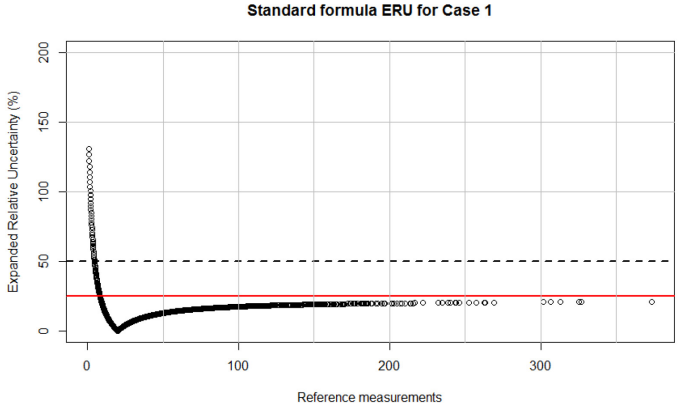


Figure B.5a: *The plot of standard formula REU in % for Case 1 with* $\beta_0 = 2$ *and* $\beta_1 = 0.9$.

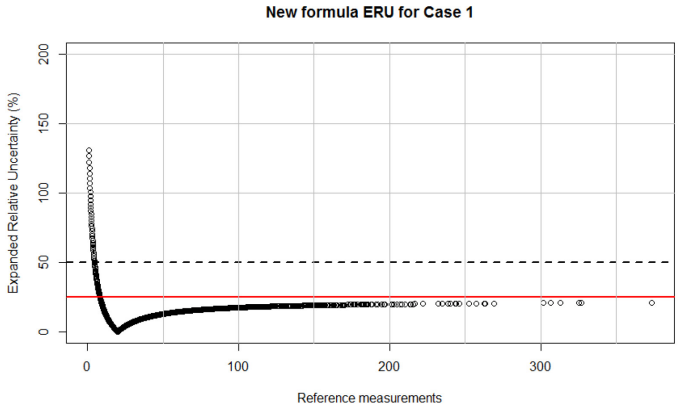


Figure B.5b: *The plot of alternative formula REU in % for Case 1 with* $\beta_0 = 2$ *and* $\beta_1 = 0.9$.

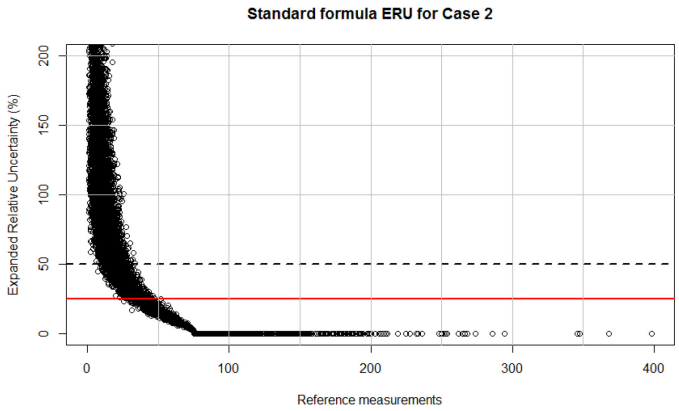


Figure B.6a: *The plot of standard formula REU in % for Case 2 with* $\beta_0 = 2$ *and* $\beta_1 = 0.9$.

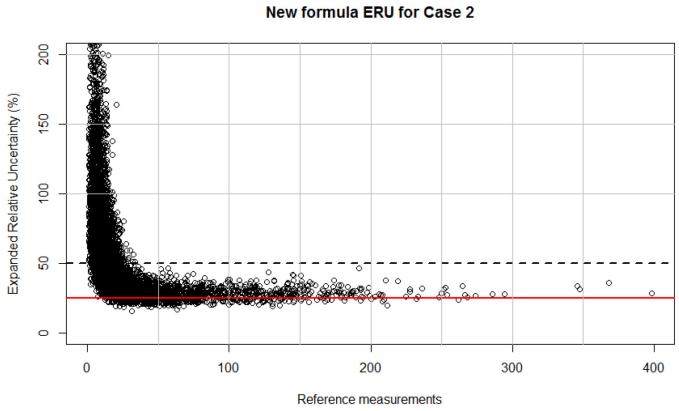


Figure B.6b: *The plot of alternative formula REU in % for Case 2 with* $\beta_0 = 2$ *and* $\beta_1 = 0.9$.

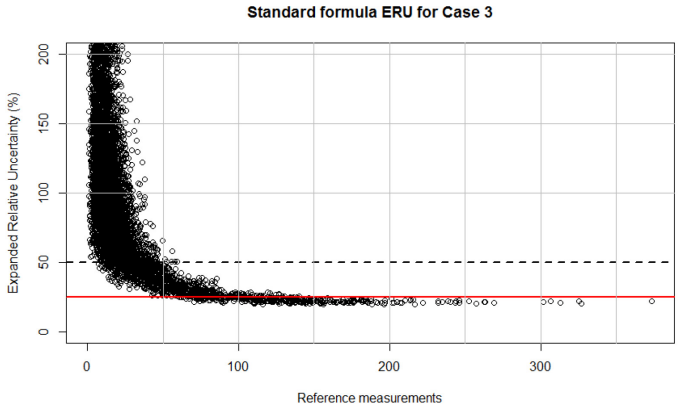


Figure B.7a: *The plot of standard formula REU in % for Case 3 with* $\beta_0 = 2$ *and* $\beta_1 = 0.9$.

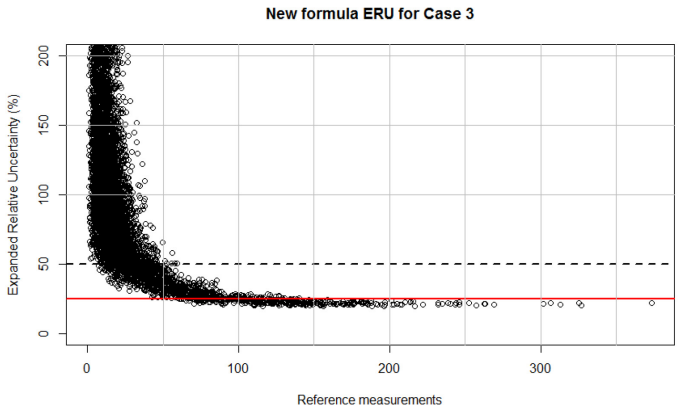


Figure B.7b: *The plot of alternative formula REU in % for Case 3 with* $\beta_0 = 2$ *and* $\beta_1 = 0.9$.

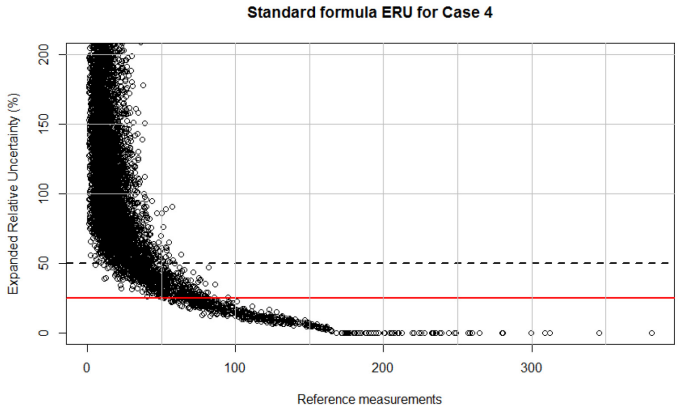


Figure B.8a: *The plot of standard formula REU in % for Case 4 with* $\beta_0 = 2$ *and* $\beta_1 = 0.9$.

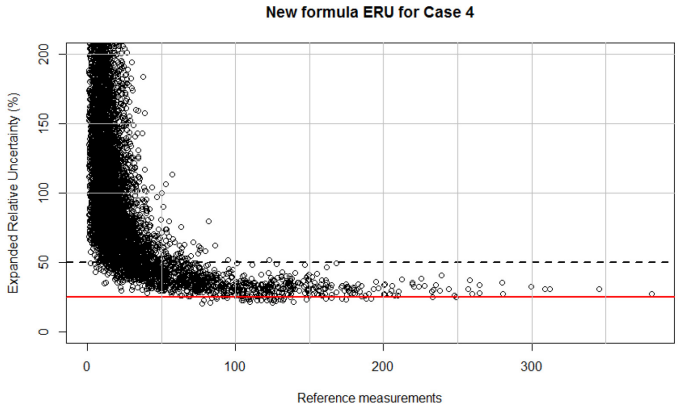


Figure B.8b: *The plot of alternative formula REU in % for Case 4 with* $\beta_0 = 2$ *and* $\beta_1 = 0.9$.

Appendix C

The `reu` R package

An easy-to-use R package called `reu` for user-friendly calculation of the relative expanded uncertainty formula has been developed as part of this project. The package consists of two main functions: `calc_reu()` and `plot_reu()`.

The function `calc_reu()` takes as inputs vectors of the reference concentrations and the sensor observations, as well some information about the pollutant and whether the "standard" or "alternative" method should be used as described in Sections 4 and 5 in this report. It produces a vector of relative expanded uncertainties in percent as output. A short description of the inputs and outputs of this routine is shown in Figure C.1.

The function `plot_reu()` creates a standardized plot of relative expanded uncertainty against the reference concentrations. The plot further includes a Loess fit to the data and provides indicators for the Data Quality Objective (DQO) and a given limit value of the concentration. An example of the standard output that the function provides is given in Figure C.2. The plots provided by the package are based on the `ggplot2` framework and are quite expandable. The function returns a `ggplot` object which can be further modified after figure production. The internal R documentation of the plotting function `plot_reu()` is shown in Figure C.3.

One limitation of the package is currently the lack of choice regarding how the reference uncertainty is calculated. More work will be necessary to offer different options for this.

calc_reu {reu} R Documentation

Calculates the relative expanded uncertainty (REU)

Description

Calculates the relative expanded uncertainty (REU)

Usage

```
calc_reu(
  reference_observations,
  sensor_observations,
  species,
  method = "standard"
)
```

Arguments

`reference_observations`	Vector with reference observations
`sensor_observations`	Vector with sensor observations
`species`	String describing the pollutant (e.g. 'no2'). This is used for picking default values of reference uncertainty.
`method`	String determining the calculation method, either 'standard' (default) or 'alternative' (see report XXX for details)

Value

Vector with relative expanded uncertainties in percent.

Figure C.1: Short documentation of the `calc_reu()` function in the R package `reu`.

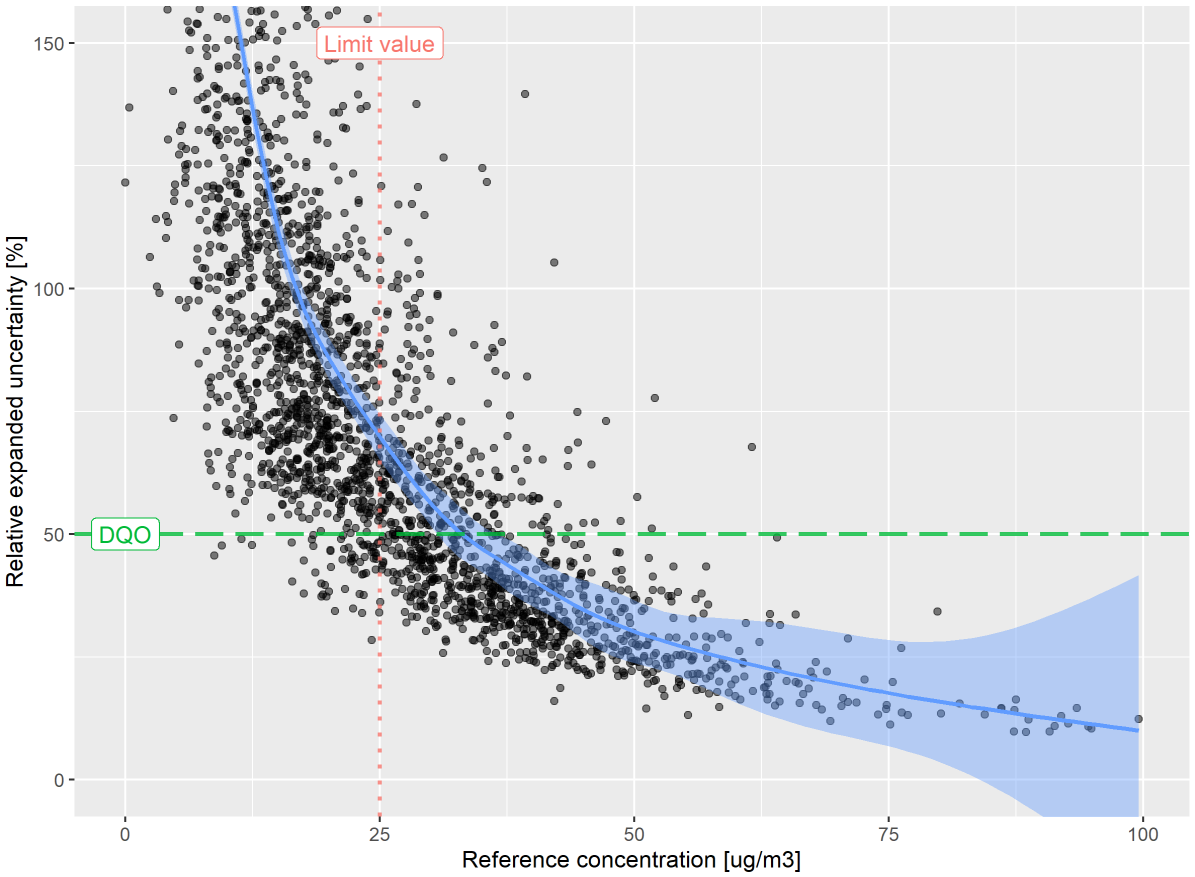


Figure C.2: Example of a standard plot by the `plot_reu()` function in the R package `reu`.

plot_reu {reu} R Documentation

Plotting function for showing the relative expanded uncertainty

Description

Plotting function for showing the relative expanded uncertainty

Usage

```
plot_reu(
  reference_concentration,
  relative_expanded_uncertainty,
  dqo = 50,
  limit_value = 25,
  ylim = c(0, 150)
)
```

Arguments

`reference_concentration`	Vector with reference concentrations
`relative_expanded_uncertainty`	Vector with relative exapnded uncertainties (output of `calc_reu()`)
`dqo`	Value of the Data Quality Objective in percent (default dqo=50)
`limit_value`	The limit value (in concentration units) for the given species. Used to draw a vertical line. E.g. 25 (default = 25)
`ylim`	Vector of two values determining the y axis limits (default c(0,200))

Value

A ggplot object with the plot

Figure C.3: Short documentation of the `plot_reu()` function in the R package `reu`.

NILU – Norwegian Institute for Air Research

NILU – Norwegian Institute for Air Research is an independent, nonprofit institution established in 1969. Through its research NILU increases the understanding of climate change, of the composition of the atmosphere, of air quality and of hazardous substances. Based on its research, NILU markets integrated services and products within analysing, monitoring and consulting. NILU is concerned with increasing public awareness about climate change and environmental pollution.

NILU's values: *Integrity - Competence - Benefit to society*
NILU's vision: *Research for a clean atmosphere*

NILU – Norwegian Institute for Air Research
P.O. Box 100, NO-2027 KJELLER, Norway

E-mail: nilu@nilu.no
http://www.nilu.no

ISBN: 978-82-425-2997-8
ISSN: 2464-3327

Norsk institutt for luftforskning
Norwegian Institute for Air Research